基于 FCBF 特征选择和集成优化学习的基因表达数据分类算法 *

马　超

(深圳信息职业技术学院 数字媒体学院，广东 深圳 518172)

摘　要：针对微阵列基因表达数据高维小样本、高冗余且高噪声的问题，提出一种 FCBF 特征选择和集成优化学习的分类算法 FICS-EKELM。首先使用快速关联过滤方法 FCBF 滤除部分不相关特征和噪声，找出与类别相关性较高的特征集合；其次，运用抽样技术生成多个样本子集，在每个训练子集上利用改进乌鸦搜索算法同步实现最优特征子集选择和核极限学习机 KELM 分类器参数优化，然后基于基分类器构建集成分类模型对目标数据进行分类识别，此外运用多核平台多线程并行方式进一步提高算法计算效率。在六组基因数据集上的实验结果表明，不仅能用较少特征基因达到较优的分类效果，并且分类结果显著高于已有和相似方法，是一种有效的高维数据分类方法。

关键词：特征选择；集成学习；微阵列基因表达数据；乌鸦搜索算法；核极限学习机

中图分类号：TP183　　doi: 10.3969/j.issn.1001-3695.2018.04.0248

Gene expression data classification based on FCBF feature selection and ensemble optimized learning method

Ma Chao

(*College of Digital Media*, *Shenzhen Institute of Information Technology*, *Shenzhen Guangdong* 518172, *China*)

Abstract: In order to solve the problems of microarray gene expression data with the characteristic of high dimension and small sample, high redundancy and a lot of noise, this article proposed a novel model FICS-EKELM, which was build based on the combination FCBF feature selection and ensemble optimized method, for gene expression data classification. In the proposed method, Fast Correlation-based Filter method(FCBF) firstly used to eliminate the irrelevant features and noise, and chose the discriminate feature subsets. Secondly, bootstrap technology produced many sample training subsets, by means of these subsets, the improved crow search algorithm(ICS) used to select optimal feature subsets and parameters for kernel extreme learning machine(KELM) synchronously. And then, ensemble classifiers were constructed for target gene data classification, which based on the basic classifiers. Moreover, the model implemented in parallel on multi-core processor, which used OpenMP to speed up the search and optimization process. Experiment on six public famous gene datasets, the proposed method not only achieves a higher classification performance with less characteristic genes, but also greatly improves the classification accuracy. It proves the effective and validity of the proposed method.

Key words: feature selection; ensemble learning; microarray gene expression data; crow search algorithm; kernel extreme learning machine

0　引言

DNA 微阵列技术是生物信息学领域中一个具有重大意义的技术性突破，已被广泛应用于药物研究、基因测序等多个领域。通过微阵列技术得到的基因表达数据称为微阵列基因表达数据，通常表示为矩阵形式，分析的是基因发生的改变，基因间的互相关系以及基因活动产生的影响。分类是微阵列基因表达数据挖掘的一个重要任务，通过分析微阵列基因表达数据可为对疾病诊断和治疗提供可靠的分类结果。但微阵列基因表达数据中存在维度高样本少的“维数灾难”问题，导致传统模式分类研究难以解决[1]，如何进行有效特征选择和分类以识别出少部分对分类最有贡献的基因，提高分类效果，成为基因表达数据分类研究的关键问题之一。

近些年，对基因表达数据的系统性分析已成为人工智能领域中的热门研究课题[2-4]。目前有很多数据降维方法被用于基因数据的特征选择与识别。其中基因特征选择方法是基因表达数据降维最主要的方法，依据是否独立于后续的学习算法，可分为过滤式(Filter)和封装式(Wrapper) [5]：

a) Filter 法。Filter 特征选择方法一般使用评价准则来增强特征与类的相关性，削减特征之间的相关性，例如信息增益 IG、

收稿日期：2018-04-12；**修回日期：**2018-06-05　　**基金项目：**国家自然科学基金青年基金资助项目（61303113）；广东省自然科学基金资助项目（2016A030310072）；深圳市科技计划项目（GJHZ20150316112246318，KJYY20170724152553858）；校级科研课题（QN201716）

作者简介：马超（1983-），男，辽宁葫芦岛人，讲师，博士，主要研究方向为数据挖掘、机器学习、模式识别（billmach@163.com）．

最小冗余最大相关 mRMR、ReliefF、FCBF、Fisher Score 评分和最小平方回归误差等[6]。Filter 与后续学习算法无关，一般直接利用所有训练数据的统计性能评估特征，但 Filter 方法未考虑特征与分类器之间的相关性，并不能保证选择出一个优化特征子集，即使能找到一个满足条件的优化子集，它的规模也会比较庞大，会包含一些明显的噪声特征，评估与后续学习算法的性能偏差较大。

b) Wrapper 法。Wrapper 方法作为学习算法组成部分，直接以分类器的分类性能作为特征重要性程度的评价标准，跟据选择的特征子集构造最终的分类模型。这类方法是特征选择研究领域的热点，相关研究工作也较多，虽然在运行速度要比 Filter 法慢，但所选择的特征子集规模相对要小得多，有利于特征的辨识，分类准确率较高。

针对上述两种方法特点，目前很多研究采用的都是 Filter 与 Wrapper 的混合方法。例如 2016 年 Jerzy 等人[7]利用 ReliefF，mRMR 结合 SVM 方法对高维癌症数据进行分类，得到较好的分类结果；同年谢娟英等人[8]为解决基因特征选择难题，提出一种基于 K-S 检验与 mRMR 原则的混合方法，并以 SVM 为分类器，以 F1_measure、分类精度和 AUC 作为评价指标进行基因选择评估，结果证明了该方法有效性；Lai 等人[9]将 Filter 和 Wrapper 方法结合提出信息增益 IG 和改进简化群算法 ISSO 并利用线性核 SVM 分类器进行基因特征选择；2017 年 Lu 等人[10]提出结合最大化交互信息 MIM 和自适应遗传算法的混合特征选择算法来降低基因表达数据维度；同年 Wang 等人[11]针对基因表达数据维数灾难问题提出加权离散细菌优化算法进行特征基因选择，结果证明该方法能解决传统细菌优化算法过早收敛的问题；Chen 等人[12]采用粗糙集和熵计算的方法用于基于选择，结果显示该方法能够有效提高肿瘤数据的分类精度；Wang 等人[13]引入马尔可夫毯以改进 Wrapper 方法进行基因特征选择，结果证明了该方法的有效性；2018 年吴辰文等人[14]提出一种基于 ReliefF 和蚁群算法的特征基因选择方法以解决微阵列数据多分类问题，实验结果证明该方法能以较少特征基因得到较高的多分类效果；Jain 等人[15]针对基因分类和癌症诊断问题提出整合相关特征选择 CFS 和改进二元粒子群 iBPSO 算法，并利用朴素贝叶斯分类器进行分类，取得了较高的分类精度。

从这些研究可以发现，Filter 和 Wrapper 结合的方法在基因数据分类中取得了很好的效果，但仍存在两个主要问题：

a)大多数算法采用的都是 SVM 分类器，SVM 典型难题是模型的参数选择问题，这对分类结果有重要影响，但对参数选择没有统一的标准和理论指导；

b)现有方法都采用单一分类器模型，关于集成学习的基因特征分类方法研究很少，基因数据分类的精度可能会由于单一分类器性能而达到瓶颈。而核极限学习机(kernel extreme learning machine，KELM)具有比 SVM 和 BP 神经网络(BPNN)更优的性能[16]。基于上述分析，为了克服上述不足，得到准确率更高的分类效模型，本文提出了一种新颖的分类模型 FICS-EKELM，用于高维基因表达数据的分类研究。

本方法首先利用 FCBF 特征选择进行初步特征选择，剔除掉数据集中冗余特征及噪音；然后使用 bootstrap 抽样进行 PCA 转换生成多个训练样本子集，在每个训练子集上采用改进乌鸦搜索算法 ICS 同步进行最优特征子集选择和 KELM 模型参数优化，得到具有差异度的基 KELM 分类器，最后构建集成 KELM 分类模型，并运用多核平台多线程并行方式进一步提高算法运算效率，最后对测试集进行测试。

本文工作创新点如下：

a)运用 FCBF 方法对原始高维基因表达数据进行特征降维，剔除掉数据集中冗余特征及噪音，与其它 Filter 方法相比，FCBF 算法既考虑了特征间相关性又分析了特征的冗余性，并且在大量实验比较中，FCBF 被证明具有较低的时间复杂度和较好的特征选择结果，而 ReliefF、IG 以及 Fisher Score 等方法虽然能处理不完备和有噪音的数据，但未能很好地处理冗余特征。

b)提出 ICS 算法同步进行特征子集选择和 KELM 参数优化，构建基分类器。乌鸦搜索算法 CSA 简单易实现、涉及参数较少，与粒子群算法 PSO、遗传算法 GA 以及人工蜂群算法 ABC 等算法相比，都能得到近似甚至更优的优化结果。

c)采用集成分类思想进行基因特征选择和分类。

d)基于多核处理技术运用 OpenMP 来实现模型并行运算，可以有效提高算法的效率。

1　理论介绍

1.1　FCBF 算法(fast correlation-based filter)

基于快速关联的过滤算法 FCBF[17]是一种典型的启发式序列后向消除方法，使用对称的不确定度量来衡量两个特征的相关性。算法核心思想是采用对称不确定性(Symmetrical uncertainty, SU)作为度量标准，如果一个特征与类别之间的不确定性程度高，且与其它已选特征之间的不确定性程度低，则将该特征标记为重要特征。

FCBF 算法简单描述如下：

给定数据集(x_i, t_i)， $i=1, \dots, N$, 其中 $x_i=[x_{i1}, x_{i2}, \dots, x_{id}]^T\in R^n$， $t_i=[t_{i1}, t_{i2}, \dots, t_{im}]^T\in R^m$，样本类别为 $Y=(y_1, y_2,\dots y_N)$.

Step1:初始化 T 和 S；　/*T 为特征向量集合，S 为特征子集*/

Step2:对于每个 $t_i\in T$，计算特征与类别的 SU 值，即 $SU(t_i,Y)$，其计算公式为 $SU(A,B)=2\left[\frac{H(A)-H(A|B)}{H(A)+H(B)}\right]$；

Step3:选出 T 中 $SU(t_i,Y)>r$ 的特征，根据 SU 值降序排序并存入 S'中；

Step4:从 S'中选择一个特征 t_i，将 t_i 存入 S 集合中，并从 S'中删掉 t_i；

Step5:计算 t_i 与 t_j 的对称不确定性 SU 值 SU(t_i, t_j)，去除 t_j 的冗余特征，如果 SU$(t_i, t_j)> SU(t_i,Y)$，则从 S'中删掉 t_i；

Step6:Repeat Step4 和 Step5 until S 为空集；

Step7:输出得到的特征子集 S.

1.2　核极限学习机(KELM)

KELM 是由 Huang 等人[18]在单隐层前馈神经网络模型 ELM 基础上提出的方法，它能逼近任意连续目标函数，其输出值能以极小误差逼近类标签值。

假设给定 N 个训练样本集(x_i, t_i)， $i=1, \dots, N$, 其中 $x_i=[x_{i1}, x_{i2}, \dots, x_{id}]^T\in R^n$， $t_i=[t_{i1}, t_{i2}, \dots, t_{im}]^T\in R^m$，隐层激活函数为

$g(x)$，隐层节点个数为 L，ELM 输出函数的计算公式为：

$$f(x)=\sum_{j=1}^{L}\beta_j h_j(x)=h(x)\beta \tag{1}$$

其中 $\beta_j=[\beta_{j1}, \beta_{j2}, \dots, \beta_{jL}]^T$ ($j=1, 2, \dots, L$) 表示连接第 j 个隐层节点和输出层节点间的输出权重值。其中 $H=\{h_{ij}\}$ ($i=1, \dots, N$; $j=1, \dots, L$)为隐层输出矩阵，H 第 j 个隐层节点的第 j 列对应输入 $x_1, x_2, \dots, x_n$，H 的第 i 行对应输入 x_i 的输出向量。通常采用最小二乘法确定线性系统的输出权重值：

$$\beta'=H^T T \tag{2}$$

其中:H^+为隐层输出矩阵 H 的 Moore-Penrose 广义逆矩阵。

之后 Huang 等引入核函数避免 ELM 方法随机产生输入权重和偏倚值的问题，提出基于核函数的 ELM 方法 KELM，KELM 输出权重的计算公式如下：

$$\beta=H^T(I/C+HH^T)^{-1}T \tag{3}$$

因此，KELM 输出函数的表达式为：

$$f(x)=h\beta=h(x)H^T(I/C+HH^T)^{-1}T \tag{4}$$

当隐层映射函数 $h(x)$不可知时，核函数矩阵计算公式如下：

$$\Omega_{ELM}=HH^T:\Omega_{ELMi,j}=h(x_i)\cdot h(x_j)=K(x_i,x_j) \tag{5}$$

其中 $K(x_i, x_j)$表示核函数，KELM 中核函数为 RBF 核函数，那么 KELM 分类模型的计输出函数表达式为：

$$f(x)=\begin{bmatrix}K(x,x_1)\\ \dots\\ K(x,x_N)\end{bmatrix}^T(I/C+\Omega_{ELM})^{-1}T \tag{6}$$

2　改进乌鸦搜索算法 ICS(improved crow search algorithm)

乌鸦搜索算法 CSA 是由 Askarzadeh 于 2016 年提出的一种新的元启发算法[19]，它模拟的是自然界中乌鸦的智能觅食行为。

在求解最优问题时，假定 N 只乌鸦随机分布在 n 维搜索空间中，$x^{i,t}=[x_1^{i,t}, x_2^{i,t},\dots,x_n^{i,t}]$($i=1,2,\dots,N$; $t=1,2,\dots,Maxiter$)表示第 i 只乌鸦在第 t 次迭代时的位置。$M^{i,t}$ 表示乌鸦 i 在第 t 次迭代时隐藏食物的记忆值，即最优位置。$AP^{i,t}$ 表示乌鸦 i 在第 t 次迭代时的感知概率 AP，$fl^{i,t}$ 表示乌鸦 i 在第 t 次迭代时的飞行长度；

对乌鸦搜索算法进行初始化控制参数设置，所述初始化控制参数包括种群群体数量 M、感知概率 AP、飞行长度 fl 以及最大迭代次数 $Maxiter$；

传统乌鸦搜索算法是随机初始化位置，公式如下所示：

$$x^{i,t}=rand\cdot(x_{\max}-x_{\min})+x_{\min} \tag{7}$$

其中，$x^{i,t}$ 为乌鸦随机产生的位置，$x_{\max}$ 为 x 的最大值，$x_{\min}$ 为 x 的最小值，$rand$ 为[0,1]区间随机生成数。

但是随即初始化导致个体质量无法保证，如果初始解群较好，将会有助于求解效率与解的质量，如果不好则会影响求解效率，增加了不确定性，一个好的初始化种群能够确保算法更快地收敛，本文将混沌算法优化乌鸦搜索来解决上述问题。混沌运动是在确定性非线性系统中自然出现的类随机行为，它具有确定性过程同时也兼具随机性[20]。混沌运动可以使得算法能够跳出局部最优的同时寻找全局最优解。因此本文采用混沌映射函数 Logistics 对乌鸦位置进行初始化：

$$X_{n+1}=\mu\cdot X_n\cdot(1-X_n)\quad \mu\in[0,4], X_n\in(0,1) \tag{8}$$

其中参数 μ 用于控制混沌程度。

通过感知概率 AP 进行动态调整以达到全局搜索和局部搜索的平衡状态，由于乌鸦位置的更新影响着最优解和收敛速度，引入混沌算法进一步优化乌鸦搜索位置的更新，位置更新的表达式如下：

$$x^{i,t+1}=\begin{cases}x^{i,t}+w_i\cdot r_i\cdot fl^{i,t}\cdot(m^{j,t}-x^{i,t}), & \text{if}\ \ w_z\ge AP^{j,t}\\ rand\cdot(x_{\max}-x_{\min})+x_{\min}, & else\end{cases} \tag{9}$$

其中 w_i 表示在第 i 代时得到的混沌映射值，w_z 表示在第 z 代得到的混沌映射值，$AP^{i,t}$ 表示乌鸦 j 在 t 代时的感知概率，r_i 和 r_j 是[0,1]区间均匀分布的随机数。

由公式(9)可知，通过混合函数的引入进一步平衡算法全局搜索和局部搜索，对全局搜索和局部搜索进行更加灵活地动态扰动，在前期 w_i 值较大确保全局搜索占较大权重，提高种群搜索的多样性，到迭代后期，w_i 值变小，使得局部搜索权重加大，加速算法收敛。

当乌鸦 i 的位置发生改变，则更新记忆值表达式如下：

$$M^{i,t+1}=\begin{cases}x^{i,t+1}, & \text{if}\ f(x^{i,t+1})>f(M^{i,t})\\ M^{i,t}, & else\end{cases} \tag{10}$$

其中，$M^{i,t}$ 表示乌鸦记忆值，$f(M^{i,t})$表示适应度值。

对于二进制乌鸦搜索算法在离散空间内进行搜索，每个解表示为 1 或 0，引入映射函数 $S(x)$将连续空间的值转换到离散空间[0,1]，计算公式如下：

$$M^{i,t+1}=\begin{cases}1, & \text{if}\ f(S(M^{i,t+1}))\ge rand()\\ 0, & else\end{cases} \tag{11}$$

其中 $rand()$为[0,1]区间均匀分布的随机数。映射函数 $S(x)$ 表达式如下：

$$S(M^{i,t+1})=\frac{1}{1+e^{10(M^{i,t+1}-0.5)}} \tag{12}$$

3　FICS-EKELM 模型

本节对 FICS-EKELM 模型进行详细说明，模型整体架构如图 1 所示。

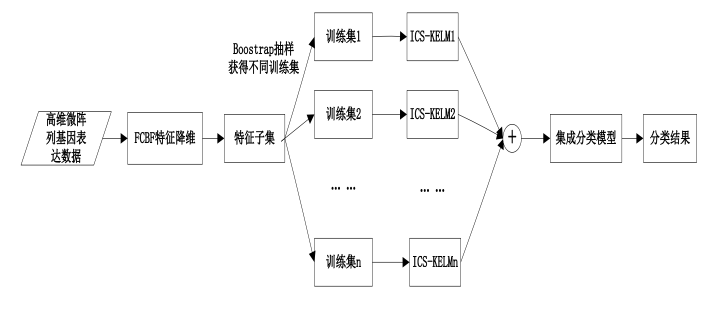


图 1　FICS-EKELM 模型的总体流程图

3.1 产生训练子集

为保证数据样本的多样性，引入旋转森林算法思想[21]，通过 boostrap 抽样方法从原始数据中随机抽取样本，并进行 PCA 转换，产生新样本集。假设原始数据集样本为 X，类标号为 Y，新的训练集中样本数为 k，产生训练样本的算法具体过程如下所示：

```
Input：Original datasets X
Output：sub_datasets(T1, T2, …, Tk)
Begin
      For i = 1 to k
         [sub_X, sub_Y] = randomsub (X);
         trainX_subnew = bootstrapal(sub_X, sub_Y);   /*进行抽样*/
         Coeff = pcasky(trainX_subnew); /*进行 PCA 转换得出新样本*/
         R_coeff = sort (Coeff);            /*进行排序*/
         New_sub (i) = trainX_subnew* R_coeff;
      End For
End
      Return：Final sub_datasets (T1, T2, …, Tk)
```

3.2 构建基分类器模型

文献[22]中明确指出：对于构建集成分类模型，要得到更高分类精度的充要条件是分类器必须是准确且存在差异的，即具有较大差异的分类器集成模型具有更强的性能，因此构建差异性的分类器是一个重要问题。与 SVM 一样，KELM 受其惩罚因子 C 和核宽 γ 的影响较大，若取值不当，会导致模型分类效果较差。构建基于 KELM 模型差异性集成分类器，可通过以下两个条件来实现：(1)通过 boostrap 采样和 PCA 特征转换可得到不同的训练数据集，使得每个 KELM 模型得到不同的输入样本，保证在不同的训练数据集上训练 KELM 模型；(2)影响 KELM 分类性能的重要参数惩罚因子 C 和核宽 γ，难以人为设置，采用 ICS 算法进行优化，能得到不同的分类模型，从而保证数据多样性和分类器差异性。

本文基分类器 ICS-KELM 的核心思想是利用 ICA 算法进化机制同步进行特征子集选择和参数优化，从而得到最优基分类器。构建基分类器模型的流程图如图 2 所示，具体步骤如下：

a)种群初始化，群体中每个个体由多个特征属性离散值，以及惩罚因子 C 和核宽 γ 两个连续值构成。编码形式为 $\tau=(1,0,\ldots,1,1,C,\gamma)$，其中 1 为选中的特征，0 为未选中特征；

b)利用初始化个体解码所得到的参数在训练子集上进行 KELM 训练，计算每个个体的适应度值，适应度值计算公式为：

$$Fitness=\alpha\cdot acc_i+(1-\alpha)\cdot\frac{N-|Subset|}{N} \tag{14}$$

其中 acc_i 表示第 i 个解的分类精度，N 为特征总数，$|Subset|$ 表示选出的最优特征子集的特征数，α 为调节分类精度和特征子集数量两部分的权重值，$0<\alpha<1$，本文中 α 取值为 0.8，*Fitness* 表示 K 折交叉验证(*K*-fold Cross Validation, K-fold CV)平均值。

c)增加迭代次数；

d)更新种群的位置和记忆值，比较种群个体的适应度值，若新的适应度值大于比较值，则将最优适应度值记为新适应度值；

e)利用上一步得到新的个体进行解码所得到的参数在 KELM 上训练，并根据公式(14)计算其适应度值；

f)如果达到最大种群数量，转到步骤 d 执行，否则转到步骤 g 执行；

g)比较当前适应度值和全局最优适应度值，若当前值大于记录的最优适应度值，更新为当前值；

h)若达到最大迭代次数，算法转到步骤 i 执行，否则转到步骤 c；

i)输出全局最优的记忆值位置，即为最优解；

j)利用得到最优特征子集和参数在训练子集上训练，构造出最优的基分类器。

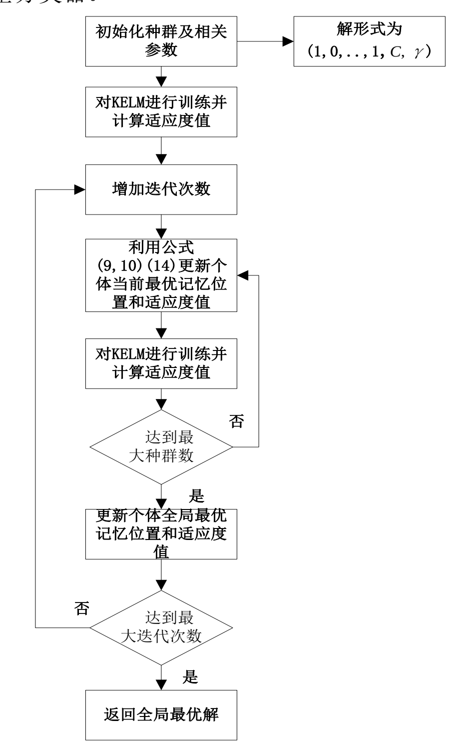


图 2 ICS 优化 KELM 构建基分类器的流程图

3.3 集成分类器模型

本文采用加权投票法将这些不同的基分类器集成为一个分类器模型，权重系数由基分类器对验证集的分类准确率归一化处理后获得，分类器组合输出最终的输出结果。对于给定的数据样本 x，模型中有 K 个分类器 $Tk(x)$，$k=1, 2, \ldots, K$，那么多数投票策略得到样本最终的分类结果计算公式如下：

$$T(x)=\arg\max\sum_{k=1}^{K}\delta_{\mathrm{sgn}(Tk(x)),y} \tag{15}$$

其中 $\delta_{i,j}=\begin{cases}0, & i\neq j\\ 1, & i=j\end{cases}$，$y\in\{-1,1\}$是分类器的输出类标号。公式(15)表示 K 个分类器 $Tk(x)$累积结果的最大值。

3.4 并行模型计算

对于复杂优化问题，ICS 算法需要多次更新才能保证找到最优解，ICS 算法的初始解生成、适应度计算、种群位置更新等在算法中比较耗时，但它们是相互独立的，所以该算法具有天然的并行性。为充分发挥 ICS 的并行性，提高算法效率，本文提出基于多核处理器利用 OpenMP[23]来实现模型并行运算，多核平台整体框架分为三层：

a)该层由一系列粒子组成，并行算法控制整个 ICS 迭代过

程，每个个体独立参与整个运算过程。

b)OpenMP 平台。该层是为保证实现并行算法的同步，同时建立和操作系统间的通讯联系，平台核心组件是调度器，能给操作系统提供作业的调度和分配。

c)多核处理器。作业在该层通过 OpenMP 被系统调用。

并行模型 FICS-EKELM 的伪代码如下：

```
initialize model parameters
train KELM;
calculate the fitness;    /*fitness 为适应度值*/
while t<max_iteration /* max_iteration  为最大迭代次数*/
   for each solution
        update position;
        update memory;
        train KELM;
        calculate the fitness;
        calculate fitness_best;
        calculate memory_best;
   end for;
        calculate fitness_global;
        calculate memory_global;
       t=t+1;
end while
```

4　实验分析

4.1　实验设置

为了评估本方法对高维微阵列基因表达数据的有效性，分别选取 Breast Cancer、CNS、Leukemia、Lung Cancer、Lymphoma 以及 Prostate 六组公共高维基因数据集。各基因数据集的信息如表 1 所示。

表 1　基因数据集信息

数据集	基因数量	样本数量	类别
Breast Cancer	24481	97	2
CNS	7129	60	2
Leukemia	7129	72	2
Lung Cancer	7129	96	2
Lymphoma	4026	62	3
Prostate	12600	102	2

文中实验在 Windows 7 操作系统上进行，Intel Core(TM) i5 处理器，主频 3.2 GHz，内存 4 GB，在 MATLAB 2014b 环境下编程实现。ELM 和 KELM 采用 MATLAB 工具箱。ICS-KELM 算法的参数如表 2 所示。

表 2　ICS-KELM 算法的参数设置

ICS 算法参数	数值
种群数量	30
最大迭代次数	80
飞行长度 *fl*	2
感知概率 *AP*	0.1

此外，ICS-EKELM 模型实验结果与 ELM、KELM、SVM 以及 BPNN 等方法进行了比较，模型的详细参数设置如下：为了公平比较，KELM 与 SVM 模型的参数设置相同，均采用网格计算方法，模型中 C 和 γ 的搜索范围为 $C\in\{2^{-11},\ldots,2^{11}\}$ 和 $\gamma\in\{2^{-11},\ldots,2^{11}\}$。ELM 和 BPNN 方法中的隐层节点个数取值通过试凑法获得，ELM 和 BPNN 隐层节点数分别为 18 和 21 个。

4.2　实验结果讨论

为了验证提出方法的有效性，实验首先给出在六组数据集上，提出的方法与四种常用方法，即分别基于 ReliefF、mRMR、IG、CFS 特征选择方法的分类结果进行对比分析，如表 3 所示。表 3 各个算法得到的分类准确率以及标准方差值。从实验结果可以看出，在这五种模型中，提出的方法取得了最高的分类准确率，而基于 ReliefF、mRMR、IG、CFS 特征选择的方法取得的分类结果明显低于本方法，例如以 Breast Cancer 数据集为例，本方法的平均分类精度达到了 92.98%，而其它四种方法分别只得到了 88.42%、85.57%、83.51%和 81.92%的平均分类准确率，同时也说明了通过特征选择和集成分类学习，能有效提高高维基因数据的分类准确率。此外，从标准方差值可见本方法的方差值较小，也证明了该方法具有良好的稳定性。

为了充分证明所提出方法特征选择的有效性，表 4 给出了五种特征选择算法在六组数据集上所选特征个数。其中，本方法所选择的特征个数最少，FCS 和 ReliefF 次之，这是由于本方法先使用 FCBF 筛选掉大量不相关特征和噪音特征后，又利用 ICS 搜索进一步优化特征子集选择，有效剔除掉了冗余度较高的特征。

表 3　五种算法分类精度比较

数据集	本文方法	ReliefF	Fisher Score	mRMR	FCS
Breast Cancer	**92.98** ±0.21	88.42 ±0.30	85.57 ±0.34	83.51 ±0.42	81.92 ±0.33
CNS	**91.87** ±0.27	90.13 ±0.44	84.34 ±0.42	88.39 ±0.44	80.96 ±0.73
Leukemia	**99.71** ±0.18	94.58 ±0.33	93.92 ±0.38	98.52 ±0.25	94.83 ±0.45
Lung Cancer	**90.79** ±0.25	85.67 ±0.31	86.91 ±0.29	67.67 ±0.54	78.16 ±0.38
Lymphoma	**100.00**	95.26 ±0.46	90.13 ±0.35	85.21 ±0.39	82.67 ±0.44
Prostate	**97.43** ±0.31	92.98 ±0.39	86.27 ±0.47	87.95 ±0.36	84.31 ±0.30

表 4　五种方法选择的特征个数比较

数据集	本文方法	ReliefF	IG	mRMR	FCS
Breast Cancer	8	28	30	28	21
CNS	9	45	32	32	29
Leukemia	4	30	34	32	33
Lung Cancer	9	31	45	44	32
Lymphoma	7	31	28	35	31
Prostate	5	26	30	34	25

为了更好地评估分类方法的性能，表 5 给出了本方法与研究常用的分类模型 SVM、ELM、BPNN 和 NB 方法进行了比较。从表中结果可知，本文方法在分类性能上要显著优于其它四种方法，这是由于本方法在进行利用 ICS 算法进行特征选择的同

时还优化了模型参数，同时构建集成分类模式也能克服单一分类器易过拟合和分类瓶颈问题，进一步提高了分类准确率。

表 5 五种基于不同分类器的方法分类结果比较

数据集	本文方法	SVM	ELM	BPNN	Naive Bayes
Breast Cancer	**92.98**	88.62	87.83	86.60	91.75
CNS	91.87	**92.33**	90.05	89.39	90.00
Leukemia	**99.71**	95.83	94.41	96.63	98.63
Lung Cancer	**90.79**	80.21	78.96	82.26	85.44
Lymphoma	**100.00**	**100.00**	98.61	97.94	98.32
Prostate	**97.43**	96.17	97.05	96.55	97.18

基分类器的个数选取对结果是有影响的，本文对分类器集成数量进行了实验性分析，由于集成分类器数量尚未形成统一指导理论，多是通过多次实验进行尝试所得，本实验集成分类器数量的取值范围设为[1,10]，从中选取出分类结果最佳所对应的参数值用于后续实验，结果如图 3 所示。从图中可以看出，从分类器个数为 1 开始，随着分类器数量的增加，分类准确率有明显的提高，当分类器个数为 5 时，达到了最高的分类准确率，之后随着个数的增加，分类性能未得到进一步提高，而是呈现波动状态，说明在集成分类器达到一定数量后，即使继续增加分类器数量并未有助于分类性能的进一步提高。

为了验证并行模型的性能，将并行模型与串行模型进行了比较。表 6 给出了并行和串行模型在六组数据集上训练时间以及分类精度的比较情况。从表中可以看到，两个模型在分类精度指标上的结果非常相近，它们的差别在于交叉验证过程数据集的随机选择造成，但串行模型所花费的实际明显高于并行模型。

图 4 给出了并行模型和串行模型在 CNS 数据集上 5 折 CV 上独立运行的运行时间比较。从图中可见，在运行时间上，串行模型所花费的 CPU 平均运算时间大约是并行模型 PHGSA-KELM 的 2.6 倍，在每一折过程中并行模型花费的时间要远低于串行模型，这表明提出的方法从并行算法获益，弥补串行算法在迭代优化过程中耗时过多，提高了算法的计算效率。

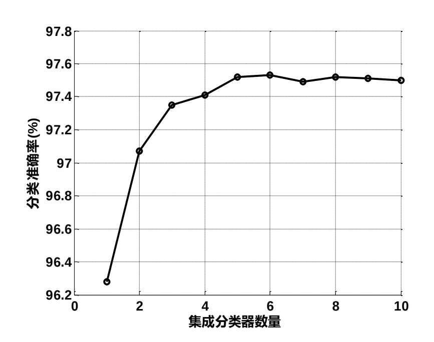


图 3 集成分类器的数量对分类性能的影响

表 6 并行模型和串行模型的训练时间和分类精度的对比

数据集	并行模型训练时间/s	串行模型训练时间/s	并行模型分类精度(%)	串行模型分类精度(%)
Breast Cancer	325.116	584.282	92.98	92.67
CNS	132.235	327.022	91.87	91.91
Leukemia	134.653	336.528	99.71	99.52
Lung Cancer	132.879	327.434	90.79	95.45
Lymphoma	85.696	212.177	100.00	100.00
Prostate	178.320	296.595	97.43	97.47

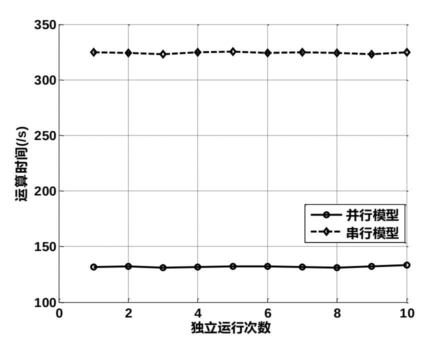


图 4 并行模型和串行模型在 5 折 CV 上的运行时间比较

为验证 ICS 算法全局搜索能力和收敛速度，实验进一步对算法的迭代机制进行研究，以 Lung Cancer 数据集为例，给出 ICS 和原始 CSA 算法在 5 折 CV 中(选的是第 1 折)的最优适应度值变化过程，如图 5 所示。图中给出的是全局最优值的变化过程，将每一次迭代中所有个体的最优适应度值记录下来。由图分析可知，性能较好的是 ICS 曲线，从第一次迭代一直到第 80 次迭代逐步演化，ICS 曲线在初始阶段增长迅速，在第 23 次迭代时收敛到最高值，之后趋于平稳；适应度值较低的是 CSA 曲线，在第 19 次迭代时收敛到较高值，之后趋于平稳，但仍低于 ICS 曲线值，说明 CSA 算法有可能陷入局部最优而未找到全局最优解。该现象证明了 ICS 算法比原始 CSA 算法具有更优的全局搜索能力和收敛速度，能迅速收敛到全局最优解。

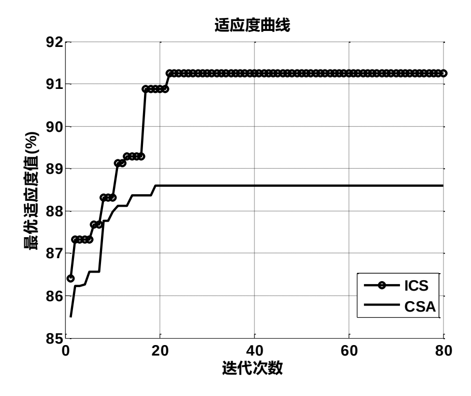


图 5 ICS 和 CSA 算法在第 1 折上训练得到最优适应度值

从图 3~5 中的结果可知，本方法在六组公共基因数据集上分别取得了 92.98%、91.87%、99.71%、90.79%、100.00%以及 97.43%的平均分类准确率，同时通过特征选择在原始高维数据特征空间下，分别得到了 8、9、4、9、7 和 5 个特征个数，极大降低了特征数量，但分类精度却没有得到明显下降，说明了特征选择的有效性。这是由于设计评估函数同时考虑了特征选择和分类器性能，在尽可能减少特征数量的同时最大化分类结果。从表 6 和图 4 中的结果对比可知，通过并行计算方式，将本方法从模型训练所花费的 CPU 平均运算时间比串行计算时间节省了近 2/3，说明通过并行计算方式能充分发挥 ICS 的并行性，提高算法效率。

5 结束语

为了更好地处理高维基因数据的分类问题，本文充分利用 FCBF 过滤性能、乌鸦搜索能力以及集成分类模型的优势，提出了一种基于 FCBF 和集成优化 KELM 分类器的分类模型。在该方法中，FCBF 方法可以有效去除数据集中冗余特征及噪声，找出对分类结果相关程度较高的特征集合，降低了特征维度，同时采用集成分类方法思想，将 KELM 作为基分类器，并利用 ICS 算法同步实现最优特征子集选择和 KELM 模型参数优化，以得到最优分类模型，且基于多核处理器利用 OpenMP 实现了模型并行运算，进一步提升了算法计算效率。实验结果表明，提出的方法优于其它典型过滤特征选择方法，以及基于 SVM、ELM、BPNN 以及 NB 的分类方法。本方法不仅可以去除冗余基因，还取得了较高的分类效果，实验验证了方法的有效性和高效性。

下一步工作将对特征与特征子集间关系的度量进行研究。此外，对 ICS 算法中设定参数的灵活调整，降低参数设置对搜索以及分类结果的影响，实现对算法的进一步优化，也是值得研究的内容。

参考文献：

[1] Boulesteix A L, Strobl C, Augustin T, *et al.* Evaluating microarray based classifiers: an overview [J]. Cancer Informatics, 2008, 6: 77–97.

[2] Lin Hungyi. Reduced gene subset selection based on discrimination power boosting for molecular classification [J]. Knowledge-Based Systems, 2018, 142: 181-191.

[3] Hanaa S, Gamal A, Nawal E. Classification of human cancer diseases by gene expression profiles [J]. Applied Soft Computing, 2017, 50: 124-134.

[4] Liang Sen, Ma Anjun, Yang Sen, *et al.* a review of matched pairs feature selection methods for gene expression data analysis [J]. Computational and Structural Biotechnology, 2018, 16: 88-97.

[5] Ghaddar B, Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines [J]. European Journal of Operational Research, 2018, 265 (3): 993-1004.

[6] Bolón-Canedo V, Sanchez-Marono N, Alonso-Betanzos A. A review of feature selection methods on synthetic data [J]. Knowledge and Information Systems, 2013, 34 (3): 483−519.

[7] Jerzy K, Tomasz L. The feature selection bias problem in relation to high-dimensional gene data [J]. Artificial Intelligence in Medicine, 2016, 66: 63-71.

[8] 谢娟英，胡秋锋，董亚非. K-S 检验与 mRMR 相结合的基因选择算法 [J]. 计算机应用研究, 2016, 33 (4): 1013-1018. (Juanying Xie, Qiufeng Hu, Yafei Dong. Gene selection algorithm based on K-S test and mRMR [J]. Application Research of Computers, 2016, 33 (4): 1013-1018.)

[9] Lai C M, Yeh W C, Chang C Y. Gene selection using information gain and improved simplified swarm optimization [J]. Neurocomputing, 2016, 218: 331-338.

[10] Lu Huijuan, Chen Junying, Yan Ke, *et al.* A hybrid eature election lgorithm for gene expression data classification [J]. Neurocomputing, 2017, 256: 56-62.

[11] Wang Hong, Jing Xingjian, Niu Ben. A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data [J]. Knowledge-Based Systems, 2017, 126: 8-19.

[12] Chen Yumin, Zhang Zunjun, Zheng Jianzhong, *et al.* Gene selection for tumor classification using neighborhood rough sets and entropy measures [J]. Journal of Biomedical Informatics, 2017, 67: 59-68.

[13] Wang Aiguo, An Ning, Yang Jing, *et al.* Wrapper-based gene selection with Markov blanket [J]. Computers in Biology and Medicine, 2017, 81: 11-23.

[14] 吴辰文，李晨阳，郭叔瑾，等. 基于 ReliefF 和蚁群算法的特征基因选择方法 [J]. 计算机应用研究, 2018, 35 (9): 1-7. (Wu Chenwen, Li Chenyang, Guo Shujin, *et al.* Feature gene selection method based on ReliefF and ant colony optimization [J]. Application Research of Computers, 2018, 35 (9): 1-7.)

[15] Jain I, Vinod K, Jain R. Correlation feature selection based improved binary particle swarm optimization for gene selection and cancer classification [J]. Applied Soft Computing, 2018, 62: 203-215.

[16] Luo Fangfang, Guo Wenzhong, Yu Yuanlong. A multi-label classification algorithm based on kernel extreme learning machine [J]. Neurocomputing, 2017, 260: 313-320.

[17] Song Q B, Ni J J, Wang G T. A fast clustering-based feature subset selection algorithm for high-dimensional data [J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25 (1): 1-14.

[18] Huang Guangbin. Extreme learning machine for regression and multiclass classification [J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42 (2): 513-529.

[19] Askarzadeh A. A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm [J]. Computers & Structures, 2016, 169: 1–12.

[20] Wang G G, Guo L, Gandomi A H, *et al.* Chaotic krill herd algorithm [J]. Information Sciences, 2014, 274: 17–34.

[21] Rodriguez J, Kuncheva L, Alonso C J. Rotation forest: a new classifier ensemble method [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28 (10): 1619-1630.

[22] Hansen L K, Salamon P. Neural network ensembles [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1990, 12 (10): 993-1001.

[23] Chapman B, Jost G, Van R. Using OpenMP: portable shared memory parallel programming [M]. Cambridge: MIT Press, 2007.